Bounds on the Power-weighted Mean Nearest Neighbor Distance

By Elia Liitiäinen, Amaury Lendasse and Francesco Corona

Department of Computer Science and Engineering, Helsinki University of Technology,
P.O. Box 5400, 02015 Espoo, Finland
elia.liitiainen@hut.fi

In this paper, bounds on the mean power-weighted nearest neighbor distance are derived. Previous work concentrates mainly on the infinite sample limit, whereas our bounds hold for any sample size. The results are expected to be of importance for example in statistical physics, nonparametric statistics and computational geometry, where they are related to the structure of matter as well as properties of statistical estimators and random graphs.

Keywords: nearest neighbor distances, bounds, random graphs

1. Introduction

Let $(X_i)_{i=1}^M$ be a set of random variables sampled from a probability distribution P on $\Re^n$ with a density q on a bounded set $\mathcal{C}$ (with respect to the Lebesgue measure λ). The k-th nearest neighbor distance of the point X_i in the l^p-norm is denoted by $d_{i,k}$. In this paper we examine the expected power-weighted distance $E[d_{i,k}^\alpha]$ with $\alpha > 0$. This quantity plays a significant role in many fields including nonparametric statistics (see Kohler *et al.* (2006) and Evans & Jones (2002)), physics (Torquato (1995)) and geometric probability (Penrose & Yukich (2003)). Nearest neighbor distances have turned out to be important in such tasks as convergence analysis of statistical estimators and theoretical analysis of particle systems in statistical physics; furthermore, nearest neighbor graphs are a fundamental class of graphs in computational geometry.

A large part of the previous work on the topic concentrates on the asymptotic form of $M^{\alpha/n}E[d_{i,k}^\alpha]$; for example, denoting by $V_{n,p}$ the volume of the unit ball, defining $\Gamma(\cdot)$ as the Gamma function and taking $p = 2$, it is shown in Evans *et al.* (2002) that

$$M^{\alpha/n}E[d_{i,k}^\alpha] \to V_{n,2}^{-\alpha/n}\frac{\Gamma(k+\alpha/n)}{\Gamma(k)}\int_{\mathcal{C}} q(x)^{1-\alpha/n}dx \tag{1.1}$$

as $M \to \infty$ assuming that q is smooth and positive on a convex and compact set $\mathcal{C}$. Moreover, similar asymptotic results can also be derived for more general random structures than the nearest neighbor graph, see for example Penrose & Yukich (2003).

In this paper, instead of the forementioned asymptotic analysis, we derive non-asymptotic lower and upper bounds for $E[d_{i,k}^\alpha]$. Denoting by $B_p(0,r)$ the open ball of radius r and center at the origin, we assume that $\mathcal{C} \subset B_p(0,\sqrt{n}/2)$ to

ensure that all nearest neighbor distances are smaller than $\sqrt{n}$. In the special case $\mathcal{C} = [-1/2, 1/2]^n$, our main result can be summarized as follows (with the notation $\|\cdot\|_p$ for the function L^p-norm w.r.t. λ):

Theorem 1.1. *The inequalities*

$$\delta_{M,k,\alpha} = \frac{1}{M}\sum_{i=1}^{M} d_{i,k}^{\alpha} \leq (\frac{2^n k}{V_{n,p} M})^{\alpha/n} + o(M^{-\alpha/n}) \leq (\frac{n^{n/2} 2^n k}{M})^{\alpha/n} \tag{1.2}$$

hold almost surely when $0 < \alpha < n$. *The latter inequality is valid also for* $\alpha = n$.

A remarkable fact about Theorem 1.1 is its universality as the upper bounds hold for (almost) any combination of points in $\mathcal{C}$ being based on purely deterministic arguments.

Previous upper bounds on the average power-weighted k-nearest neighbor distances include Kohler *et al.* (2006), Kulkarni & Posner (1995) and Torquato (1995). Compared to our bounds, the proof technique used in Kulkarni & Posner (1995) gives much higher constants when α is close to n; actually for $\alpha = n$, the bound contains an additional logarithmic factor. The probabilistic upper bounds in Kohler *et al.* (2006) on the other hand do not yield the explicit form of the constants, while Torquato (2005) concentrates on hard sphere systems.

In the special case $k = 1$, an essentially similar upper bound as ours has been derived by Tewari & Gokhale (2004). However, the bound is not rigorously derived as analysis of the boundary effect is excluded. While the authors discuss a physical model where taking the limit $M \to \infty$ is realistic, for a moderate sample size the points close to the boundary tend to have a significant effect and should be taken into account.

Our lower bound differs from (1.2) by being based on a measure theoretic argument. To our knowledge, there is no previous work on non-asymptotic lower bounds.

Theorem 1.2. *The inequality*

$$E[d_{i,k}^{\alpha}] \geq 3^{-\alpha/2} 2^{-\alpha/n} e^{-\alpha/2n} \|q\|_2^{-2\alpha/n} V_{n,p}^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}$$

holds for $\alpha > 0$ *and* $1 \leq i \leq M$.

It is worth noticing that by Lemma 5.1 in Evans *et al.* (2002),

$$\frac{\Gamma(M)}{\Gamma(M+\alpha/n)} = M^{-\alpha/n}(1 + O(\frac{1}{M})) \text{ as } M \to \infty.$$

2. A Geometric Upper Bound

The formal definition of the nearest neighbor of the point X_i in the l^p-norm is

$$N[i,1] = \text{argmin}_{1 \leq j \leq M, j \neq i} \|X_i - X_j\|_p.$$

The k-th nearest neighbor is defined recursively as

$$N[i,k] = \text{argmin}_{1 \leq j \leq M, j \neq i, N[i,1], \ldots, N[i,k-1]} \|X_i - X_j\|_p,$$

that is, the closest point after removal of the preceeding neighbors. The corresponding distances are defined as

$$d_{i,k} = \|X_i - X_{N[i,k]}\|_p.$$

Notice that the definition of $N[i,k]$ is not necessarily unique if there are two points at the same distance from X_i. However, the probability of such an event is zero and consequently it can be be neglected as Theorem 1.1 is stated to hold almost surely. Moreover, the conclusion of Theorem 1.1 holds for all configurations of points in $\mathcal{C}$ with an arbitrary method of tie-breaking. For example, $N[i,k]$ can be chosen as the smallest index among the alternatives.

To fix some notation, let us define

$$\mathcal{C}_r = \{x \in \Re^n : \exists y \in \mathcal{C} \text{ s.t. } \|x - y\|_p \leq r\}.$$

Next we prove a geometric upper bound for the average k-nearest neighbor distance. The proof is based on showing that any point $x \in \mathcal{C}$ belongs to at most k balls $B_p(X_i, d_{i,k}/2)$.

Theorem 2.1. *For any $0 < \alpha \leq n$ and $r > 0$,*

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k}^{\alpha} I(d_{i,k} \leq r) \leq \left(\frac{2^n k \lambda(\mathcal{C}_{r/2})}{V_{n,p} M}\right)^{\alpha/n} \tag{2.1}$$

almost surely.

Proof. Choose any $x \in \Re^n$. Let us make the counterassumption that there exists $k+1$ points, denoted by $X_{i_1}, \ldots, X_{i_{k+1}}$ (the indices being distinct), such that $x \in B_p(X_{i_j}, d_{i_j,k}/2)$ for $j = 1, \ldots, k+1$. Let $(i_j, i_{j'})$ be the pair that maximizes the distance $\|X_{i_j} - X_{i_{j'}}\|_p$. Under these conditions the triangle inequality yields

$$\|X_{i_j} - X_{i_{j'}}\|_p < \frac{1}{2} d_{i_j,k} + \frac{1}{2} d_{i_{j'},k}.$$

The strict inequality holds because $B_p(x,r)$ is an open ball. On the other hand,

$$\begin{aligned}
\|X_{i_j} - X_{i_{j'}}\|_p &= \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\|_p + \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\|_p \\
&= \frac{1}{2}\max_{1 \leq j' \leq k+1} \|X_{i_j} - X_{i_{j'}}\|_p + \frac{1}{2}\max_{1 \leq j \leq k+1} \|X_{i_j} - X_{i_{j'}}\|_p \\
&\geq \frac{1}{2} d_{i_j,k} + \frac{1}{2} d_{i_{j'},k}
\end{aligned}$$

leading to a contradiction. Thus we have for the sum of indicator functions

$$\begin{aligned}
&\sum_{i=1}^{M} \int_{\Re^n} I(x \in B_p(X_i, d_{i,k}/2), d_{i,k} \leq r) dx \\
&= \sum_{i=1}^{M} \int_{\mathcal{C}_{r/2}} I(x \in B_p(X_i, d_{i,k}/2), d_{i,k} \leq r) dx \\
&\leq \lambda(\mathcal{C}_{r/2}) k.
\end{aligned}$$

On the other hand,

$$\sum_{i=1}^{M} \int_{\Re^n} I(x \in B_p(X_i, d_{i,k}/2), d_{i,k} \leq r)dx = 2^{-n} V_{n,p} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r)$$

implies that

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r) \leq 2^n k V_{n,p}^{-1} \lambda(\mathcal{C}_{r/2}) M^{-1}.$$

By Jensen's inequality,

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^\alpha I(d_{i,k} \leq r) \leq (\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r))^{\alpha/n},$$

which implies Equation (2.1). □

As stated in the following corollary, the last inequality in Theorem 1.1 follows straightforwardly by choosing $r = \sqrt{n}$ because $\mathcal{C} \subset B_p(0, \sqrt{n}/2)$ implies that all nearest neighbor distances are smaller than $\sqrt{n}$ and $\lambda(\mathcal{C}_{\sqrt{n}/2}) \leq V_{n,p} n^{n/2}$.

Corollary 2.2. *For $0 < \alpha \leq n$ we have*

$$\delta_{M,k,\alpha} \leq (\frac{2^n k n^{n/2}}{M})^{\alpha/n}.$$

3. The Boundary Effect

In this section we show that the bound in Theorem 2.1 can be improved by dividing the nearest neighbor distances into two different sets corresponding to small and large values. We will show that the volume of the set $\cup_{i=1}^{M} B_p(X_i, d_{i,k}/2) \setminus \mathcal{C}$ is asymptotically neglible, which consequently implies the first inequality in (1.2).

Theorem 3.1. *Assume that $\lambda(\mathcal{C}_r) \leq \lambda(\mathcal{C}) + c_1 r$ when $r \leq c_2$ for some constants $c_1, c_2 > 0$. Then for any $0 < \alpha < n$,*

$$\begin{aligned} \frac{1}{M} \sum_{i=1}^{M} d_{i,k}^\alpha &\leq \sup_{0 \leq r \leq \sqrt{n}} (2^\alpha k^{\alpha/n} \lambda(\mathcal{C}_{r/2})^{\alpha/n} V_{n,p}^{-\alpha/n} M^{-\alpha/n} + 2^n n^{n/2} k r^{\alpha-n} M^{-1}) \\ &= 2^\alpha k^{\alpha/n} \lambda(\mathcal{C})^{\alpha/n} V_{n,p}^{-\alpha/n} M^{-\alpha/n} + O(M^{-\frac{\alpha n - \alpha^2 + n}{n^2 - \alpha n + n}}). \end{aligned} \quad (3.1)$$

Proof. Let us define the set of indices

$$I_r = \{1 \leq i \leq M : \ d_{i,k} > r\}$$

corresponding to points with the k-th nearest neighbor distance larger than r. If the number of elements $|I_r|$ is bigger than one, we may define the subsample $(X_i)_{i \in I_i}$

and the corresponding nearest neighbor distances d_{i,k,I_r}. Because excluding points can only increase the distances between a point and its nearest neighbors, we obtain

$$\frac{1}{M}\sum_{i=1}^{M} I(d_{i,k} > r) d_{i,k}^{\alpha} \leq \frac{1}{M}\sum_{i\in I_r} d_{i,k,I_r}^{\alpha}.$$

A straightforward application of theorem 2.1 yields

$$\frac{1}{M}\sum_{i\in I_r} d_{i,k,I_r}^{\alpha} \leq 2^{\alpha} k^{\alpha/n} n^{\alpha/2} |I_r|^{1-\alpha/n} M^{-1}.$$

The first inequality in (3.1) follows now by Chebyshev's inequality:

$$|I_r| = \sum_{i=1}^{M} I(d_{i,k} > r) \leq 2^n r^{-n} n^{n/2} k.$$

One should also take in the account the case $|I_r| = 1$ which, however, does not pose any problems as $r \leq \sqrt{n}$.

To see the second result, choose

$$r = M^{-\frac{n-\alpha}{n^2-\alpha n+n}}$$

and use the approximation $(1+x)^{\alpha/n} \approx 1 + \alpha n^{-1} x$ valid for small x. □

The condition $\lambda(\mathcal{C}_r) \leq \lambda(\mathcal{C}) + c_1 r$ requires some regularity of the boundary of $\mathcal{C}$. It is similar to condition C.2 in Evans *et al.* (2002). Such a bound holds for most sets encountered in practice; for example, if $\mathcal{C} = [-1/2, 1/2]^n$ we have

$$\lambda(\mathcal{C}_{r/2}) - \lambda(\mathcal{C}) \leq (1+r)^n - 1 = nr + O(r^2).$$

It is clear that the influence of points close to the boundary grows once the dimensionality of the space becomes bigger. To demonstrate the improvement obtained compared to the straight application of theorem 2.1 with $r = \sqrt{n}$, both bounds are plotted in figure 1 for $n = 3$, $k = 1$, $p = 2$, $\alpha = 1$ and $\mathcal{C} = [-1/2, 1/2]^3$ using the estimate $\lambda(\mathcal{C}_{r/2}) \leq (r+1)^3$ in (3.1).

4. A Probabilistic Lower Bound

(*a*) *The Small Ball Probability*

In this section the lower bound in theorem 1.2 is derived. The proof is based on the properties of the random variable

$$\omega_{X_i}(d_{i,k}) = P(B_p(X_i, d_{i,k})).$$

It is interesting that $\omega_{X_i}(d_{i,k})$ has a distribution that is independent of the probability measure P as shown in the following well-known lemma. The proof is given here for completeness, but it can also be found for example in Evans & Jones (2002).

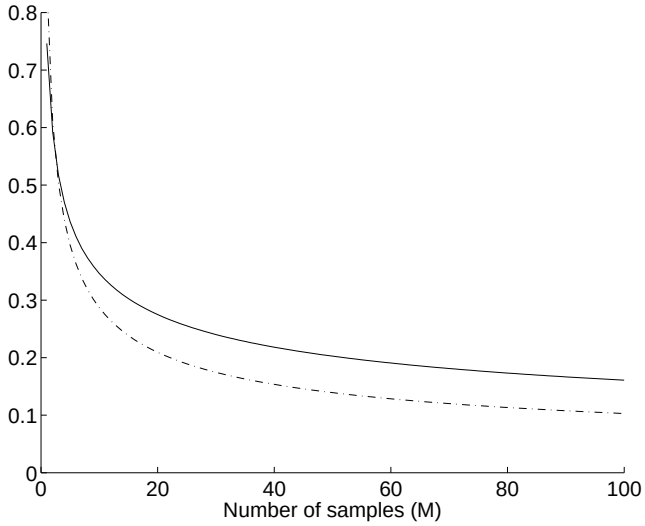

Figure 1. A demonstration of the bounds in Corollary 2.2 (the solid line) and Theorem 3.1.

Lemma 4.1. *For any* $i, \alpha > 0$,

$$E[\omega_{X_i}(d_{i,k})^\alpha | X_i] = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)}.$$

Proof. Choose $0 < z < 1$ and set $t = \inf\{s > 0 : \ \omega_{X_i}(s) > z\}$. Then $\omega_{X_i}(t) = z$ and $\omega_{X_i}(d_{i,k}) > z$ if and only if there are at most $k-1$ points in the set $B_p(X_i, t)$. Thus a combinatorial argument yields almost surely

$$\begin{aligned} P(\omega_{X_i}(d_{i,k}) > z | X_i) &= \sum_{j=0}^{k-1} \binom{M-1}{j} \omega_{X_i}(t)^j (1-\omega_{X_i}(t))^{M-j-1} \\ &= \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1-z)^{M-j-1}. \end{aligned} \tag{4.1}$$

Uxsing the formula

$$\binom{M-1}{j} = \frac{\Gamma(M)}{\Gamma(M-j)\Gamma(j+1)},$$

Theorem 8.16 in Rudin (1986) and, for example, an induction argument together

with the properties of the Beta function we obtain

$$\begin{aligned} E[\omega_{X_i}(d_{i,k})^\alpha | X_i] &= \alpha \int_0^1 z^{\alpha-1} P(\omega_{X_i}(d_{i,k}) > z | X_i) dz \\ &= \alpha \sum_{j=0}^{k-1} \int_0^1 \binom{M-1}{j} z^{j+\alpha-1} (1-z)^{M-j-1} dz \\ &= \alpha \sum_{j=0}^{k-1} \binom{M-1}{j} \frac{\Gamma(j+\alpha)\Gamma(M-j)}{\Gamma(M+\alpha)} \\ &= \alpha \sum_{j=0}^{k-1} \frac{\Gamma(M)\Gamma(j+\alpha)}{\Gamma(j+1)\Gamma(M+\alpha)} = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)}. \end{aligned}$$

□

(*b*) *A Derivation of the Lower Bound*

Let us define the Hardy-Littlewood maximal function

$$L(x) = \sup_{t>0} \frac{\omega_x(t)}{V_{n,p} t^n}.$$

As a consequence of basic properties of maximal functions, we may bound the L^1 norm of $L(x)$ by

Lemma 4.2. *Choose $s > 1$ and let $s' = \frac{s}{s-1}$. For any $i > 0$,*

$$E[L(X_i)] \leq 3^{n/s} e^{1/s} (\frac{s^2}{s-1})^{1/s} \|q\|_s \|q\|_{s'}. \tag{4.2}$$

Proof. By Holder's inequality,

$$E[L(X_i)] \leq \|L\|_s \|q\|_{s'}. \tag{4.3}$$

By a classical result for maximal functions, see for example theorem 8.18 in Rudin (1986), $\|L(x)\|_s$ is finite if the density q belongs to the space L^s. In fact, the proof of Theorem 8.18 in the aforementioned reference gives us the well-known bound

$$\|L\|_s \leq 3^{n/s} e^{1/s} (\frac{s^2}{s-1})^{1/s} \|q\|_s,$$

which can be substituted into Equation (4.3) to obtain Equation (4.2). □

Given lemmas 4.1 and 4.2, the main result of the section follows rather easily:

Theorem 4.3. *For $\alpha, i > 0$ and $s > 1$,*

$$E[d_{i,k}^\alpha] \geq 3^{-\alpha/s} e^{-\alpha/ns} (\frac{s^2}{s-1})^{-\alpha/ns} \|q\|_s^{-\alpha/n} \|q\|_{s'}^{-\alpha/n} V_{n,p}^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}.$$

Proof. By lemma 4.1,

$$\begin{aligned} E[d_{i,k}^{\alpha}|X_i] &\geq V_{n,p}^{-\alpha/n} L(X_i)^{-\alpha/n} E[\omega_{X_i}(d_{i,k})^{\alpha/n}|X_i] \\ &= V_{n,p}^{-\alpha/n} L(X_i)^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}. \end{aligned}$$

The proof is completed by observing the fact that by Jensen's inequality

$$E[L(X_i)^{-\alpha/n}] \geq E[L(X_i)]^{-\alpha/n}$$

and applying lemma 4.2. □

Theorem 1.2 follows now as a corollary.

Corollary 4.4. *The inequality*

$$E[d_{i,k}] \geq 3^{-\alpha/2} 2^{-\alpha/n} e^{-\alpha/2n} \|q\|_2^{-2\alpha/n} V_{n,p}^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}$$

holds for $\alpha > 0$ *and* $1 \leq i \leq M$.

5. Discussion

(*a*) *Possible Improvements*

A comparison between Theorem 1.1 and Equation (1.1) reveals that our upper bound is approximately 2^{α} times higher than the asymptotic moments in (1.1). An interesting question is, whether one can actually improve theorem 3.1 by taking into account that the samples are independent and identically distributed. On the other hand, the geometric bounds are strong in the sense that they hold for any combination of points.

Theorem 4.3 requires that

$$\int_{\Re^n} q(x)^s dx < \infty$$

for some $s \geq 2$. Possibly such a condition could be avoided to obtain a similar result with weaker conditions. Another direction of considerable interest is the extension to the non-i.i.d. case, which occurs especially in physics.

(*b*) *Applications*

In general, nearest neighbor distances arise as a basic quantity in many fields. One specific application of particular interest is analyzing finite spherical packings, where the centers of non-overlapping hard spheres are chosen in a random way. Such packings arise for example via random sequential adsorption (RSA) (see Talbot *et al.* (2000)) or in equilibrium statistical physics (see for example Torquato (1995)). The main challenge in hard sphere systems is the possibility of long-range dependencies, which hamper the theoretical analysis. Thus Theorem 1.1 is an interesting tool for analysis of packing fractions and other important quantities, as it is based on a non-probabilistic argument and holds for any configuration of points. It is also

of interest to ask, whether an analogue of Theorem 1.2 could be proven for hard-sphere systems and how tight such bounds would be. The potential of bounds on nearest neighbor distances has already been noted by Torquato (1995), who found rather deep connections between physical and geometric quantities.

Many estimators in nonparametric statistics are based on the use of nearest neighbor distances. Thus nonparametric statistics is another interesting application area for the theory of nearest neighbor distributions. For some recent work we refer to Kohler *et al.* (2002), where a probabilistic nearest neighbor bound plays an important role in the convergence analysis of nonparametric regressions estimators for unbounded data. Apart from regression, similar techniques are useful in the analysis of nonparametric classifiers as demonstrated by Kulkarni & Posner (1995). The advantage of our bounds is that they take a concrete form without unknown constants while being rather tight and general.

6. References

Evans, D. & Jones A. 2002 A proof of the Gamma test. *Proc. R. Soc. Lond. A* 458, 2759-2799.

Evans, D. & Jones, A. J. & Schmidt, W. M. 2002 Asymptotic moments of near neighbour distance distributions. *Proc. R. Soc. Lond. A* 458, 2839-2849.

Kohler, M. & Krzyzak, A. & Walk, H. 2006 Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *J. Multivariate A.* 97(2), 311-323.

Kulkarni, S. R. & Posner, S. E. 1995 Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE T. Inform. Theory*, 41(4), 1028-1039.

Penrose, M.E. & Yukich, J.E. 2003 Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.* 13(1), 277-303.

Rudin, W. 1986 *Real and Complex Analysis.* Higher Mathematics Series.

Tewari, A. & Gokhale, A. M. 2004 A geometric upper bound on the mean first nearest neigbour distance between particles in three-dimensional microstructures. *Acta Mater.* 52(17), 5165-5168.

Talbot, J. & Tarjus, G. & Van Tassel, P. R. & Viot, P. 2000 From car parking to protein adsoption: an overview of sequential adsorption processes. *Colloids Surf. A* 165, 287-324.

Torquato, S. 1994 Nearest-neighbor statistics for packings of hard spheres and disks. *Phys. Rev. E*, 51, 3170-3182.

Torquato, S. 1995 Mean nearest-neighbor distance in random packings of hard d-dimensional spheres. *Phys. Rev. E*, 74 (12), 2156-2159.